PCT

WELTORGANISATION FÜR GEISTIGES EIGENTUM
Internationales Büro

INTERNATIONALE ANMELDUNG VERÖFFENTLICHT NACH DEM VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS (PCT)

(51) Internationale Patentklassifikation 6 :
G06F 17/60, 17/30, G06K 9/20

A1

(11) Internationale Veröffentlichungsnummer: WO 97/38382

(43) Internationales Veröffentlichungsdatum: 16. Oktober 1997 (16.10.97)

(21) Internationales Aktenzeichen: PCT/DE97/00583

(22) Internationales Anmeldedatum: 21. März 1997 (21.03.97)

(30) Prioritätsdaten:
196 13 400.5 3. April 1996 (03.04.96) DE

(71) Anmelder *(für alle Bestimmungsstaaten ausser US)*: SIEMENS AKTIENGESELLSCHAFT [DE/DE]; Wittelsbacherplatz 2, D-80333 München (DE).

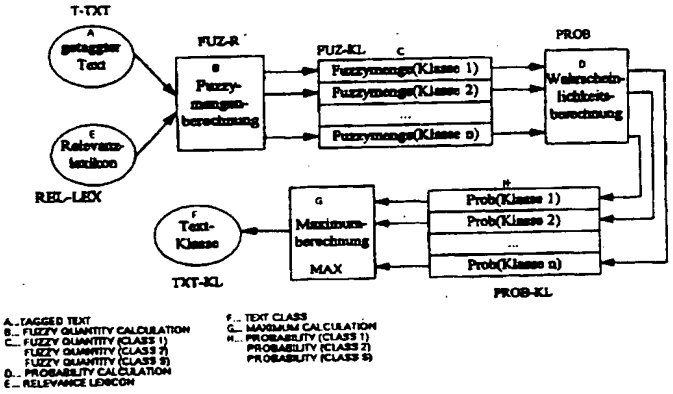
(72) Erfinder; und
(75) Erfinder/Anmelder *(nur für US)*: BLOCK, Hans-Ulrich [DE/DE]; Kirchenstrasse 42, D-81675 München (DE). BRÜCKNER, Thomas [DE/DE]; Herzogstandstrasse 24, D-81539 München (DE).

(81) Bestimmungsstaaten: JP, US, europäisches Patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Veröffentlicht
Mit internationalem Recherchenbericht.

(54) Title: METHOD OF AUTOMATICALLY CLASSIFYING A TEXT APPEARING IN A DOCUMENT WHEN SAID TEXT HAS BEEN CONVERTED INTO DIGITAL DATA

(54) Bezeichnung: VERFAHREN ZUR AUTOMATISCHEN KLASSIFIKATION EINES AUF EINEM DOKUMENT AUFGEBRACHTEN TEXTES NACH DESSEN TRANSFORMATION IN DIGITALE DATEN



(57) Abstract

The text to be classified is compared with the contents of a relevance lexicon in which the significant words of the texts to be classified are stored according to text class and their relevance for the text classes. The blurred quantity (fuzzy quantity) which indicates the occurrence per text class of the significant words of the text to be classified and their relevance for the text class is calculated. A probability calculation determines the degree of probability with which the fuzzy quantity occurs per class for the class in question. The class with the highest degree of probability is selected and the text is assigned to this class.

(57) Zusammenfassung

Der zu klassifizierende Text wird mit dem Inhalt eines Relevanzlexikons verglichen, in dem die signifikanten Wörter der zu klassifizierenden Texte pro Textklasse und deren Relevanz für die Textklassen gespeichert ist. Es wird die unscharfe Menge (Fuzzymenge) berechnet, die für die signifikanten Worte des zu klassifizierenden Textes deren Auftreten pro Textklasse und deren Relevanz für die Textklasse angibt. Mit einer Wahrscheinlichkeitsberechnung wird ermittelt, mit welcher Wahrscheinlichkeit die Fuzzymenge pro Klasse für die entsprechende Klasse auftritt. Die Klasse mit der höchsten Wahrscheinlichkeit wird ausgewählt und dieser Klasse der Text zugeordnet.

EPSDOCID: WO 9738382A1 |>

Beschreibung

Verfahren zur automatischen Klassifikation eines auf einem Dokument aufgebrachten Textes nach dessen Transformation in digitale Daten

Aus [1] ist ein System bekannt, mit dem z. B. Geschäftsbriefdokumente kategorisiert werden können und dann in elektronischer oder Papierform weitergeleitet werden können, bzw. gezielt abgelegt werden können. Dazu enthält das System eine Einheit zur Layoutsegmentierung des Dokumentes, eine Einheit zur optischen Texterkennung, eine Einheit zur Adressenerkennung und eine Einheit zur Inhaltsanalyse und Kategorisierung. Für die Segmentierung des Dokumentes wird ein gemischter bottom-up- und top-down-Ansatz benutzt, der als Einzelschritte die

- Erkennung der zusammenhängenden Komponenten,
- Erkennung der Textlinien,
- Erkennung der Buchstabensegmente,
- Erkennung der Wortsegmente und
- Erkennung der Absatzsegmente umfaßt.

Die optische Texterkennung ist in drei Teile gegliedert:

- Buchstabenerkennung in Kombination mit lexikonbasierter Wortverifikation,
- Worterkennung, mit der Klassifizierung aus Buchstaben und wortbasierter Erkennung.

Die Adresserkennung wird mit einem unifikationsbasierten Parser durchgeführt, der mit einer attributierten kontextfreien Grammatik für Adressen arbeitet. Im Sinne der Adreßgrammatik korrekt geparste Textteile sind dementsprechend Adressen. Die Inhalte der Adressen werden über Merkmalsgleichungen der Grammatik bestimmt. Das Verfahren wird in [2] beschrieben.

Für die Inhaltsanalyse und Kategorisierung werden Information-Retrieval Techniken zur automatischen Indexierung von Texten benutzt. Im einzelnen sieht dies wie folgt aus:

- Morphologische Analyse der Wörter
- Eliminierung von Stoppwörtern
- Erstellung einer Wortstatistik
- Berechnung des Indextermgewichts mit aus dem Informations-Retrieval bekannten Formeln, wie z. B. der inversen Dokumenthäufigkeit.

Mittels der so berechneten Indextermgewichte wird nun für alle Kategorien eine dreistufige Liste signifikanter Wörter ermittelt, welche die jeweilige Kategorie charakterisiert. Wie in [1] beschrieben, werden diese Listen nach der Trainingsphase noch manuell überarbeitet.

Die Kategorisierung eines neuen Geschäftsbriefes erfolgt dann durch den Vergleich der Indexterme dieses Briefes mit den Listen der signifikanten Wörter für alle Kategorien. Die Gewichte der im Brief enthaltenen Indexterme werden je nach Signifikanz mit einer Konstanten multipliziert und aufsummiert. Durch Teilen dieser Summe durch die Anzahl der Indexterme im Brief ergibt sich somit für jede Klasse eine Wahrscheinlichkeit. Die genauen Berechnungen ergeben sich aus [3]. Ergebnis der Inhaltsanalyse ist dann eine nach Wahrscheinlichkeiten sortierte Hypothesenliste.

Die der Erfindung zugrundeliegende Aufgabe besteht darin, ein Verfahren anzugeben, nach dem die Inhaltsanalyse des Textes und damit die Textklassifikation verbessert wird. Dabei wird davon ausgegangen, daß der Text des Dokumentes bereits als digitale Daten vorliegt, die dann weiterverarbeitet werden.

Diese Aufgabe wird gemäß den Merkmalen des Patentanspruches 1 gelöst.

Weiterbildungen der Erfindung ergeben sich aus den abhängigen Ansprüchen.

Ein Anwendungsfall des Verfahrens ist die automatische Diagnose aus medizinischen Befunden. Fasst man einen medizinischen Befund als Text und eine Krankheit als eine Klasse auf, so kann man das Problem der automatischen Diagnose mit dem Verfahren der Textklassifikation lösen. Ein wesentlicher Vorteil des Verfahrens ist, daß es aus einer Menge von Befunden, deren Diagnose bekannt ist, automatisch und unüberwacht das zur Klassifikation nötige Wissen lernt. Füt den Arzt ist kein zusätzlicher Aufwand nötig, er muß nur wie gewohnt den Befund schreiben. Gelernt wird aus den bereits vorhandenen Befunden. Nach der Trainingsphase wird dann mit Hilfe der gelernten Wissensquelle und Techniken der Fuzzy-Mengen ein Befund klassifiziert. Die dem Befund zugeordnete Klasse entspricht der diagnostizierten Krankheit.

Es wird zunächst davon ausgegangen, daß der zu untersuchende Text bereits in Form von ASCII-Daten vorliegt.

Vor der inhaltlichen Analyse eines Textes wird eine morphologische Analyse durchgeführt, die im ersten Schritt alle Wörter lemmatisiert (d.h. auf ihre Stammformen reduziert) und dann mit einem stochastischen Tagger lexikalische Mehrdeutigkeiten auflöst. Für die Lemmatisierung kann ein Verfahren nach [4] verwendet werden. Eine Beschreibung des verwendeten Taggers kann [5] entnommen werden. Ausgangspunkt fuer alle weiteren Bearbeitungsschritte ist immer der getaggte Text.

Die Textklassifikation ist trainingsbasiert. Aus einer Menge von Trainingstexten, deren Klassen bekannt sind, wird die Häufigkeit von Klassen, von Wörtern insgesamt und von Wörtern in den jeweiligen Klassen gezählt. Mit diesen Häufigkeiten wird dann die empirische Korrelation zwischen einem Wort und einer Klasse nach Pearson [6] berechnet. Diese Korrelation

wird für alle Wörter und alle Klassen berechnet und gilt als Relevanz eines Wortes für eine Klasse.

Berücksichtigt werden nur Korrelationen größer einem Wert r_max, der sich aus der Prüfung der Unabhängigkeit auf einem Signifikanzniveau von z. B. 0.001 ergibt (siehe hierzu z. B. [7]). Als Ergebnis erhält man ein Lexikon, das die Relevanzen der Wörter für die Klassen enthält.

Ein Text wird nach dessen morphologischer Analyse mit Hilfe dieses Relevanzlexikons wie folgt klassifiziert: Für jede Klasse wird eine unscharfe Menge ermittelt, die alle relevanten Wörtern enthält. Die Zugehörigkeitsfunktion μA der unscharfen Menge entspricht gerade dem Korrelationsmaß von Pearson. Um die wahrscheinlichste Klasse zu erhalten, wird für jede Klasse die Wahrscheinlichkeit ihrer unscharfen Menge von relevanten Wörtern berechnet. Dazu wird die in der Fuzzy-Theorie gebräuchliche Formel aus [8] benutzt, nämlich:

$$\text{prob}(A) := \sum_x \mu_A(x) \cdot p(x),$$

wobei μ_A die Zugehörigkeitsfunktion der unscharfen Menge A von relevanten Wörtern einer Klasse ist und p(x) als p(x ist relevant für A) interpretiert wird:

$$p(x \text{ ist relevant für } A) := p(A|x) = p(x,A) \;/\; p(x)$$

Als Ergebnis der Klassifikation wird die Klasse mit der wahrscheinlichsten Fuzzymenge ausgegeben.

Weiterbildungen der Erfindung ergeben sich aus den abhängigen Ansprüchen.

An Hand eines Ausführungsbeispieles wird die Erfindung weiter erläutert. Es zeigen

Figur 1 eine prinzipielle Darstellung des Verfahrens,

Figur 2 den Ablauf der Vorbereitung des Textes,

Figur 3 ein Verfahren zum Trainieren des Systems,

Figur 4 das Verfahren zur Klassifikation des Textes.

Aus Figur 1 ergibt sich eine prinzipielle Darstellung des Verfahrens. Der Text auf einem Papierdokument DOK soll klassifiziert werden. Zunächst wird das Dokument DOK mit Hilfe eines Scanners SC eingescannt und eine Bilddatei BD erzeugt. Mit Hilfe des in der europäischen Patentanmeldung 0 515 714 A1 bekannten Verfahrens wird der zu klassifizierende Text in einer Layoutsegmentierung SG segmentiert und das Textsegment TXT-SG gebildet. Man erhält wiederum eine Bilddatei, die jetzt nur noch den Textteil des Dokumentes enthält. Die Bilddaten dieses Textes werden nun mit OCR in ASCII-Daten umgewandelt. Diese sind in Fig. 1 mit TXT bezeichnet. Mit Hilfe eines Trainingslexikons REL-LEX wird die Textklassifikation TXT-K durchgeführt, somit eine Klassenhypothese erzeugt, die angibt, mit welcher Wahrscheinlichkeit der zu klassifizierende Text einer bestimmten Klasse zuzuordnen ist. Die Klassenhypothese ist in Fig. 1 mit KL-H benannt.

Vor der inhaltlichen Analyse des Textes TXT, der in ASCII-Format vorliegt, wird eine morphologische Analyse durchgeführt. Dazu werden im ersten Schritt alle Wörter des Textes lemmatisiert, d.h. auf ihre Stammformen reduziert (dies erfolgt mit Hilfe eines Lemmatisierers LEM, der den lemmatisierten Text L-TXT liefert) und dann mit einem stochastischen Tagger TAG lexikalische Mehrdeutigkeiten aufgelöst. Ergebnis dieser Behandlung des Textes TXT ist der getaggte Text T-TXT, der dann weiterverarbeitet werden kann. Die Funktionsweise des Lemmatisierers LEM ist in [4] beschrieben, der Aufbau und die Funktion des Taggers in [5].

Ausgangspunkt der weiteren Bearbeitungsschritte ist nun der getaggte Text T-TXT.

Bevor die Textklassifikation durchgeführt werden kann, muß eine Trainingsphase vorgesehen werden. In dieser Trainingsphase wird ein Relevanzlexikon REL-LEX erzeugt, das später für die Klassifikation von Texten verwendet werden wird. Dazu

wird aus einer Menge von Trainingstexten TXT-TR, deren Klassen KL-TXT bekannt sind, die Häufigkeit von Klassen, von Wörtern insgesamt und von Wörtern in den jeweiligen Klassen gezählt. Dies erfolgt in einer Einheit FR zur Frequenzberechnung, in der die Wortfrequenzen FR-W und die Klassenfrequenzen FR-KL gebildet werden. Mit diesen Häufigkeiten wird die empirische Korrelation zwischen einem Wort und einer Klasse nach Pearson [6] berechnet:

$$rlv(w \text{ in } c) := r(w,c) = \frac{N \cdot \sum wc - \sum w \cdot \sum c}{\sqrt{\left(N \cdot \sum w^2 - (\sum w)^2\right) \cdot \left(N \cdot \sum c^2 - (\sum c)^2\right)}}$$

dabei ist:
N=Anzahl der Trainingstexte,
$\sum$ wc=Anzahl der Trainingstexte der Klasse c mit Wort w,
$\sum$ w=Anzahl der Trainingstexte mit Wort w,
$\sum$ c=Anzahl der Trainingstexte der Klasse c.

Diese Korrelation wird für alle Wörter und alle Klassen berechnet und gilt als Relevanz REL eines Wortes für eine Klasse. Dabei wird beachtet, daß die Korrelationen nicht zu klein werden, es wird somit ein Wert r-max eingeführt, der z. B. auf einem Signifikanzniveau 0,001 eingestellt wird [7]. Die Ergebnisse, also die Relevanzen eines Wortes für eine Klasse werden in einem Lexikon REL-LEX abgespeichert, das also die Relevanzen der Wörter für die Klassen enthält.

Nachdem das Relevanzlexikon REL-LEX erzeugt worden ist, kann nun der zu untersuchende Text T-TXT klassifiziert werden. Dazu werden ausgewählte Wörter des Textes, die von signifikanter Bedeutung sind, aus dem Text mit den im Relevanzlexikon REL-LEX vorhandenen Beziehungen zwischen den Wörtern und den Klassen untersucht und daraus für den Text und für jede Klasse eine unscharfe Menge, eine sogenannte Fuzzymenge FUZ-R, erzeugt. Diese Fuzzymengen pro Klasse werden in einer Datei FUZ-KL abgespeichert. Die Fuzzymenge pro Klasse enthält die Worte des Textes, die in der Klasse vorkommen und deren Rele-

vanz für diese Klasse. Aus der Fuzzymenge wird für jede Klasse die Wahrscheinlichkeit ihrer unscharfen Menge von relevanten Wörtern in einer Einheit PROB berechnet und in einer Datei PROB-KL abgespeichert. Dazu wird die Zugehörigkeitsfunktion der unscharfen Menge zu der Klasse bestimmt, die gerade dem Korrelationsmaß von Pearson entspricht. Die Wahrscheinlichkeit wird nach der in der Fuzzy-Theorie gebräuchlichen Formel berechnet, diese Formel ist bereits oben angegeben worden und ist aus [8] bekannt. In einer Einheit MAX zur Maximumberechnung wird die Klasse ausgewählt, für die die höchste Wahrscheinlichkeit ausgerechnet worden ist. Dieser wird der Text T-TXT zugeordnet. Diese Klasse ist in Figur 4 mit TXT-KL benannt.

Das folgende Anwendungsbeispiel soll das Verfahren erläutern:

News aus der USENET-Newsgruppe de.comp.os.linux.misc sollen in die Klassen Drucker, Konfiguration, Netzwerk, Sound, Externer Speicher, Video, Software, Entwicklung, Kernel, Kommunikation, Eingabegeräte, SCSI, X-Windows und Betriebssystem einsortiert werden.
Der erste Bearbeitungsschritt eines Textes ist die morphologische Analyse. Sie transformiert z.B. den Satz *Beim Starten von X kommt mit der Mirage-P32 nur ein weißer Bildschirm* in die lemmatisierte Form:

```
0 1 Beim beim prp
1 2 starten starten vfin
1 2 starten starten vinfin
2 3 von von prp
3 4 X x n
4 5 kommt kommen vfin
5 6 mit mit prp
5 6 mit mit vprt
6 7 der d pron
6 7 der der det
6 7 der der relpron
```

```
7 8 Mirage mirage n
8 9 - - -
9 10 P32 p32 n
10 11 nur nur adv
11 12 ein ein det
11 12 ein ein vprt
12 13 weisser weiss adjflk
13 14 Bildschirm bildschirm n
13 15 Bildschirm. bildschirm. $$$
13 14 Bildschirm. bildschirm. $$$
14 15 . . eos_punkt
14 15 . . punkt
15 16 $CR$ $CR $CR$
```

Der Tagger löst die Mehrdeutigkeiten bei Kategorien und Grundformen auf:

```
0 1 Beim beim prp
1 2 starten starten vfin
2 3 von von prp
3 4 X x n
4 5 kommt kommen vfin
5 6 mit mit prp
6 7 der der det
7 8 Mirage mirage n
8 9 - - -
9 10 P32 p32 n
10 11 nur nur adv
11 12 ein ein det
12 13 weisser weiss adjflk
13 14 Bildschirm bildschirm n
14 15 . . eos_punkt
```

Im Training wurde folgendes Relevanzlexikon trainiert (Ausschnitt):

```
soundkarte_n
```

```
        <konfiguration> rlv = 0.012523
        <netzwerk> rlv = -0.033766
        <sound> rlv = 0.716692
        <externer speicher> rlv = -0.005260

monitor_n
        <video> rlv = 0.606806

drucker_n
        <drucker> rlv = 0.683538
        <software> rlv = 0.014210

gcc_n
        <entwicklung> rlv = 0.684036
        <kernel> rlv = 0.103325
        <kommunikation> rlv = -0.083844

apsfilter_n
        <drucker> rlv = 0.561354

grafikkarte_n
        <eingabegeraete> rlv = -0.008924
        <konfiguration> rlv = 0.017783
        <scsi> rlv = -0.005854
        <video> rlv = 0.501108

xdm_n
        <eingabegeraete> rlv = 0.023704
        <x-winows> rlv = 0.580419

scsi_n
        <eingabegeraete> rlv = -0.065260
        <kernel> rlv = -0.026075
        <konfiguration> rlv = 0.117458
        <netzwerk> rlv = -0.035671
        <betriebssystem> rlv = -0.063972
        <scsi> rlv = 0.582414
```

<sound> rlv = -0.041297
<externer speicher> rlv = 0.284832
<video> rlv = -0.107000

ethernet_n
<kommunikation> rlv = -0.012769
<netzwerk> rlv = 0.502532
<betriebssystem> rlv = 0.014134

x_n
<drucker> rlv = -0.073611
<eingabegeraete> rlv = 0.005764
<entwicklung> rlv = 0.073568
<kernel> rlv = 0.005127
<kommunikation> rlv = -0.108931
<konfiguration> rlv = -0.055763
<netzwerk> rlv = -0.077721
<betriebssystem> rlv = -0.046266
<scsi> rlv = -0.054152
<sound> rlv = -0.037581
<externe speicher> rlv = -0.081716
<software> rlv = 0.037474
<video> rlv = 0.197814
<x-winows> rlv = 0.299126

mirage_n
<scsi> rlv = 0.065466
<video> rlv = 0.221600

bildschirm_n
<drucker> rlv = -0.023347
<eingabegeraete> rlv = 0.036846
<entwicklung> rlv = -0.022288
<konfiguration> rlv = -0.014284
<video> rlv = 0.216536
<x-windows> rlv = 0.269369

starten_vinfin
<kommunikation> rlv = 0.002855
<konfiguration> rlv = 0.060185
<betriebssystem> rlv = 0.006041
<externe speicher> rlv = -0.001856
<x-windows> rlv = 0.260549

starten_vfin
<drucker> rlv = -0.038927
<entwicklung> rlv = -0.037790
<kernel> rlv = -0.009309
<kommunikation> rlv = -0.057605
<konfiguration> rlv = 0.035588
<netzwerk> rlv = 0.045992
<betriebssystem> rlv = -0.003344
<sound> rlv = -0.019409
<externe speicher> rlv = -0.043312
<video> rlv = 0.110620
<x-windows> rlv = 0.178526

Nun werden für die Klassen die Fuzzy-Mengen gebildet:
Video = {x(0.197814),mirage(0.221600),bildschirm(0.216536)}
X-Windows =
{starten(0.178526),x(0.299126),bildschirm(0.269369)}

Weiterhin sind bekannt die Wahrscheinlichkeiten der Wörter:

Wort	Video	X-Windows
x	0.24	0.19
mirage	0.8	
bildschirm	0.43	0.33
starten	0.24	0.21

Hieraus und aus den Zugehörigkeitsfunktionen der Wörter berechnen wir die Wahrscheinlichkeiten der Klassen:
Prob(Video) = 0.197814*0.24 + 0.221600*0.8 + 0.216536*0.43

Prob(X-Windows) = 0.178526*0.21 + 0.299126*0.19 +
0.269369*0.33
Prob(Video) = 0.3
Prob(X-Windows) = 0.18

Literaturverzeichnis

[1] A. Dengel et al., 'Office Maid- A System for Office Mail Analysis, Interpretation and Delivery', Int. Workshop on Document Analysis Systems (DAS)$), Kaiserslautern (1994), S 253-275.

[2] M. Malburg und A. Dengel, ' Address Verification in Structered Documents for Automatic Mail Delivery', Proc. JetPoste 93, First Europaen Conference on Postal Technologies, Vol. 1, Nantes, Frankreich (1993), S. 447-454.

[3] R. Hoch, ' Using IR Techniques for Text Classification in Document Analysis', Proc. of 17th International Conference on Research and Development in Information Retrieval (SIGIR94), Dublin, Irland, (1994).

[4] LDV-Forum, Band 11, Nummer 1 (1994), S. 17-29.

[5] E. Charniak, "Statistical Language Learning", MIT Press, Cambridge (1993),S. 45-56.

[6] H. Weber, 'Einführung in die Wahrscheinlichkeitsrechnung und Statistik für Ingenieure', 3. Auflage, Teubner, Stuttgart (1992) S. 193-194.

[7] H. Weber, 'Einführung in die Wahrscheinlichkeitsrechnung und Statistik für Ingenieure', 3. Auflage, Teubner, Stuttgart (1992) S. 244

[8] H. Bandemer und S. Gottwald, 'Einführung in Fuzzy-Methoden', 4.Auflage, Akademie Verlag, Berlin (19993), S. 161

Patentansprüche:

1. Verfahren zur automatischen Klassifikation eines auf einem Dokument aufgebrachten Textes nach dessen Transformation in digitale Daten mit Hilfe eines Rechners,

- bei dem jede Textklasse durch signifikante Wörter definiert ist,
- bei dem in einer Lexikondatei (REL-LEX) für jede Textklasse die signifikanten Wörter und deren Signifikanz für die Textklasse gespeichert werden,
- bei dem ein zuzuordnender Text mit allen Textklassen verglichen wird und für jede Textklasse die unscharfe Menge (Fuzzymenge) von Worten in Text und Textklasse und deren Signifikanz für die Textklasse ermittelt wird,
- bei dem aus der Fuzzymenge jeder Textklasse und deren Signifikanz für jede Textklasse die Wahrscheinlichkeit der Zurordnung des Textes zur der Textklasse ermittelt wird,
- bei dem die Textklasse mit der höchsten Wahrscheinlichkeit gewählt wird und dieser der Text zugeordnet wird.

2. Verfahren nach Anspruch 1,

- bei dem der zu klassifizierende Text vor der Inhaltsanalyse in einem Lemmatisierer (LEM) lemmatisiert wird,
- bei dem der lemmatiserte Text (L-TXT) einem stochastischem Tagger (TAG) zugeführt wird, um lexialische Mehrdeutigkeiten aufzulösen,
- und bei dem der getaggte Text (T-TXT) zur Textklassifikation verwendet wird.

3. Verfahren nach Anspruch 2,

- bei dem zur Klassifikation des Textes ein Relevanzlexikon (REL-LEX) erzeugt wird,
- bei dem dazu eine Menge von Trainingstexten, deren Klassen bekannt sind, verwendet wird,
- bei dem aus dieser Menge die Häufigkeit der Klassen, von Wörtern und von Wörtern in den jeweiligen Klassen gezählt wird,

- bei dem mit diesen Häufigkeiten eine empirische Korrelation zwischen einem Wort und einer Klasse berechnet wird,
- bei dem diese Korrelation für alle Wörter und alle Klassen berechnet wird und das Ergebnis der Berechnung als Relevanz eines Wortes für eine Klasse in einer Datei gespeichert wird, die als Relevanzdatei oder Relevanzlexikon (REL-LEX) verwendet wird.

4. Verfahren nach Anspruch 3,
bei dem die Korrelation (Relevanz)zwischen einem Wort und einer Klasse nach folgender Formel erfolgt:

$$rlv(w \text{ in } c) := r(w,c) = \frac{N \cdot \sum wc - \sum w \cdot \sum c}{\sqrt{\left(N \cdot \sum w^2 - (\sum w)^2\right) \cdot \left(N \cdot \sum c^2 - (\sum c)^2\right)}}$$

dabei ist:
N=Anzahl der Trainingstexte,
$\sum wc$=Anzahl der Trainingstexte der Klasse c mit Wort w,
$\sum w$=Anzahl der Trainingstexte mit Wort w,
$\sum c$=Anzahl der Trainingstexte der Klasse c.

5. Verfahren nach Anspruch 4,
bei dem nur Korrelationen > einem gewählten Wert r-max berücksichtigt werden, der auf einem Signifikanzniveau von z. B. 0.001 festgelegt wird.

6. Verfahren nach Anspruch 5,
- bei dem der zu untersuchende Text (T-TXT) und das Relevanzlexikon (REL-LEX) dazu verwendet wird, um für jede Klasse die unscharfe Menge (Fuzzymenge) der signifikanten Wörter pro Klasse und deren Relevanz pro Klasse zu ermitteln,
- bei dem aus der Fuzzymenge pro Klasse und deren Relevanz für jede Klasse die Wahrscheinlichkeit ihrer unscharfen Menge von relvanten Wörtern berechnet wird,

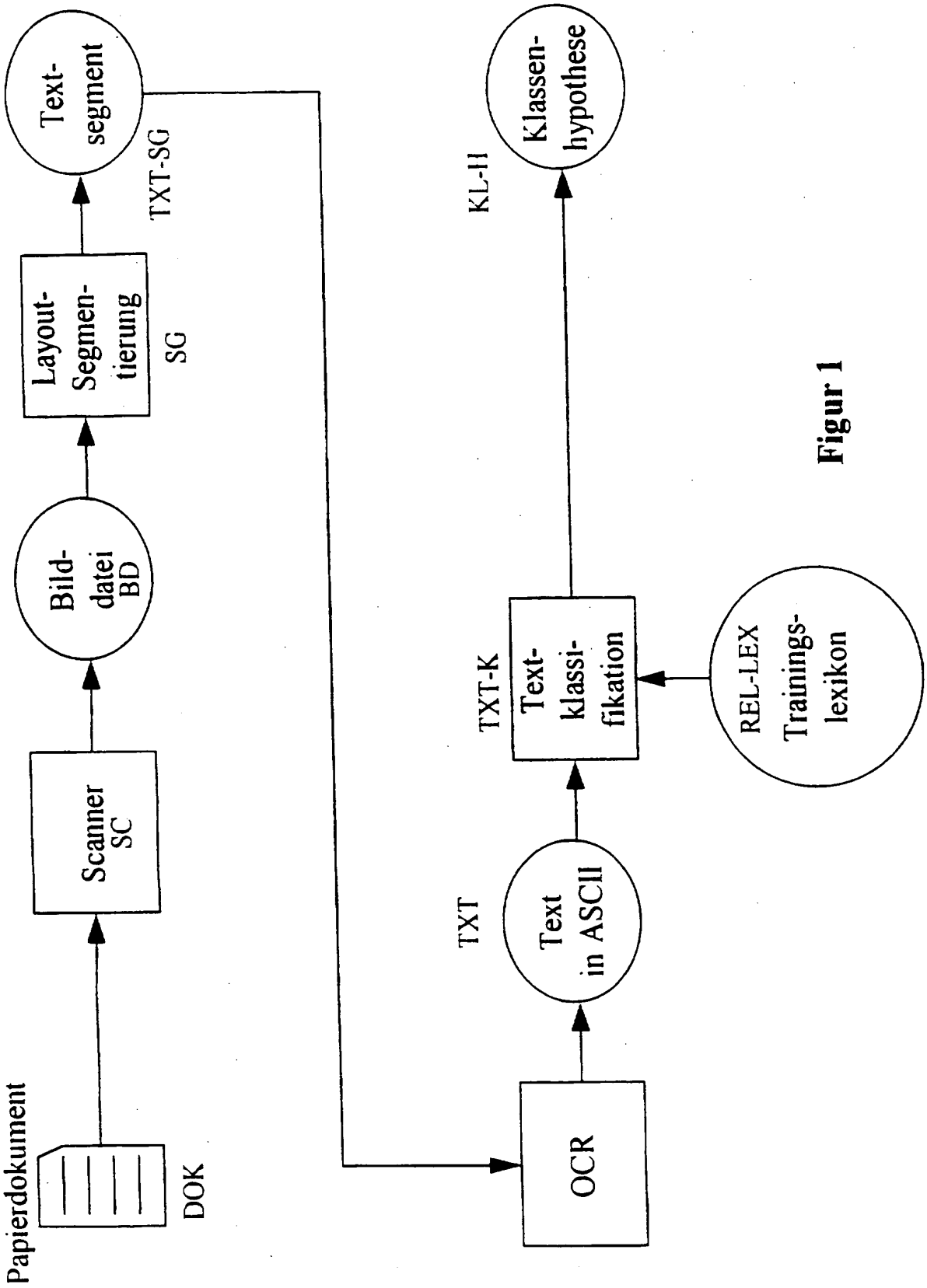
- bei dem aus den Wahrscheinlichkeiten pro Klasse die Klasse mit der maximalen Wahrscheinlichkeit ermittelt wird und dieser Klasse der Text zugeordnet wird.

7. Verfahren nach Anspruch 6,
bei dem die Berechnung der Wahrscheinlichkeit nach der Formel

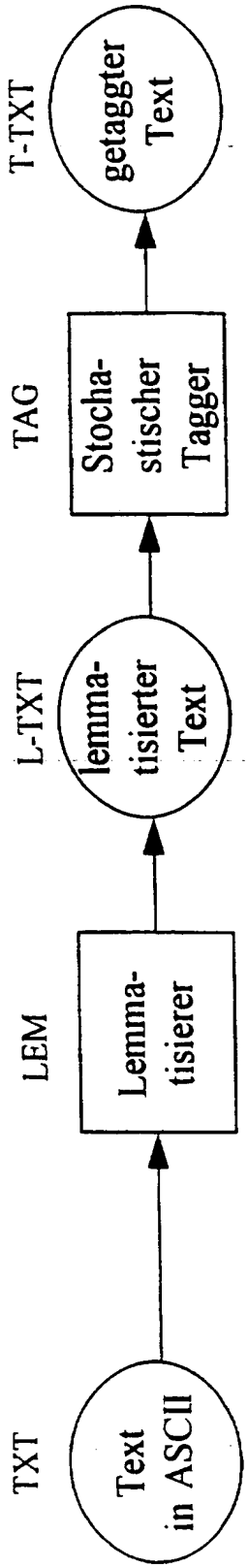
$$prob(A) := \sum_x \mu_A(x) \cdot p(x),$$

erfolgt, wobei μ_A die Zugehörigkeitsfunktion bedeutet, die angibt, in wieweit die Fuzzymenge einer Klasse zugeordnet wird und die gerade dem Korrelationsmaß nach obiger Formel entspricht.

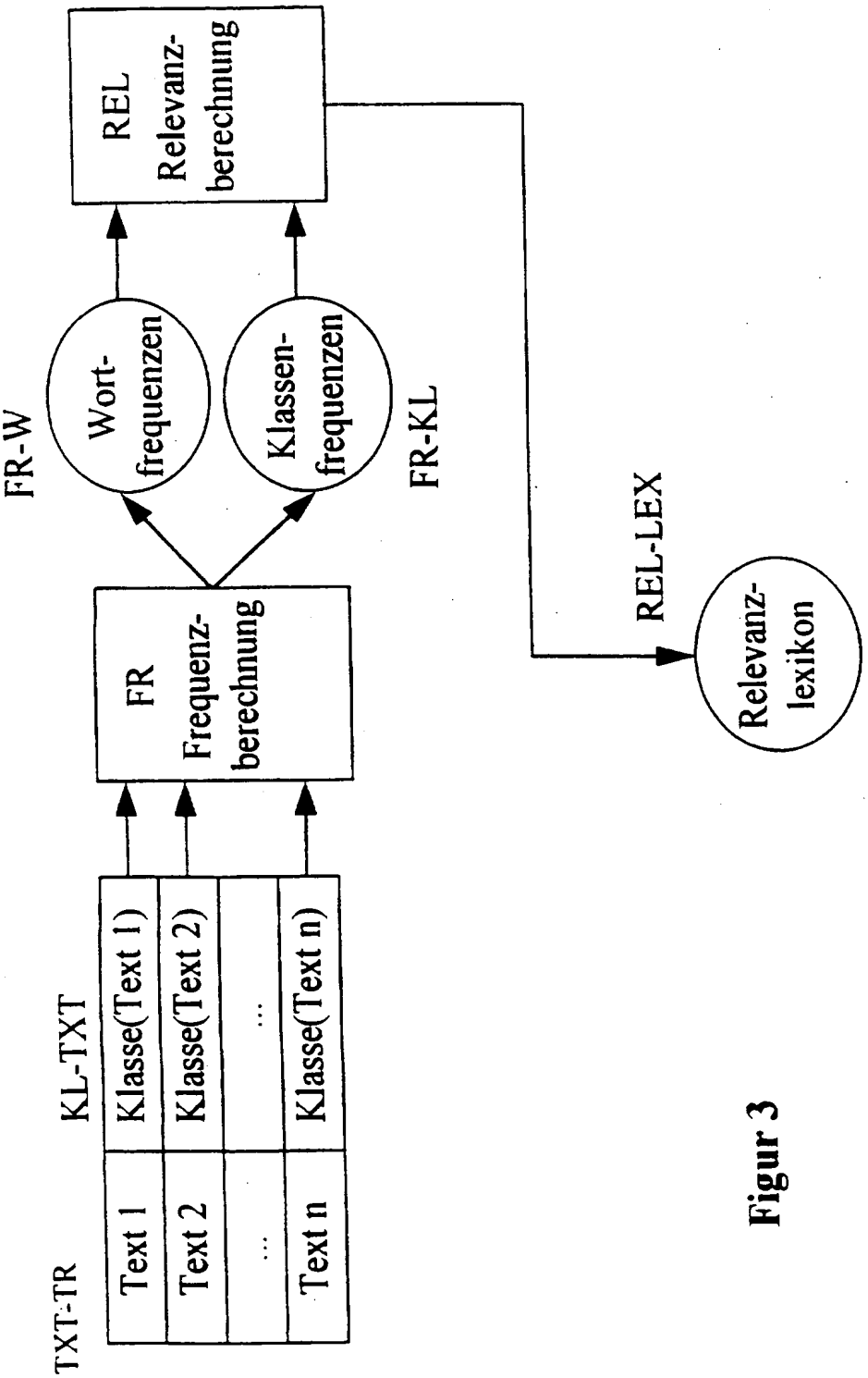
8. Verwendung des Verfahrens nach einem der vorhergehenden Ansprüche zur automatischen Diagnose aus medizinischen Befunden,
bei dem medizinische Befunde als Text und eine Krankheit als eine Klasse aufgefaßt wird
bei dem in einer Trainingsphase das zur Klassifikation erforderliche Wissen aus einer Menge von Befunden, deren Diagnose bekannt ist, automatisch gelernt wird
und bei dem ein neuer Befund nach der Technik der Fuzzymengen klassifiziert wird.


Papierdokument
DOK
Scanner SC
Bild-datei BD
Layout-Segmen-tierung
SG
Text-segment
TXT-SG
OCR
TXT
Text in ASCII
TXT-K
Text-klassi-fikation
REL-LEX Trainings-lexikon
KL-H
Klassen-hypothese

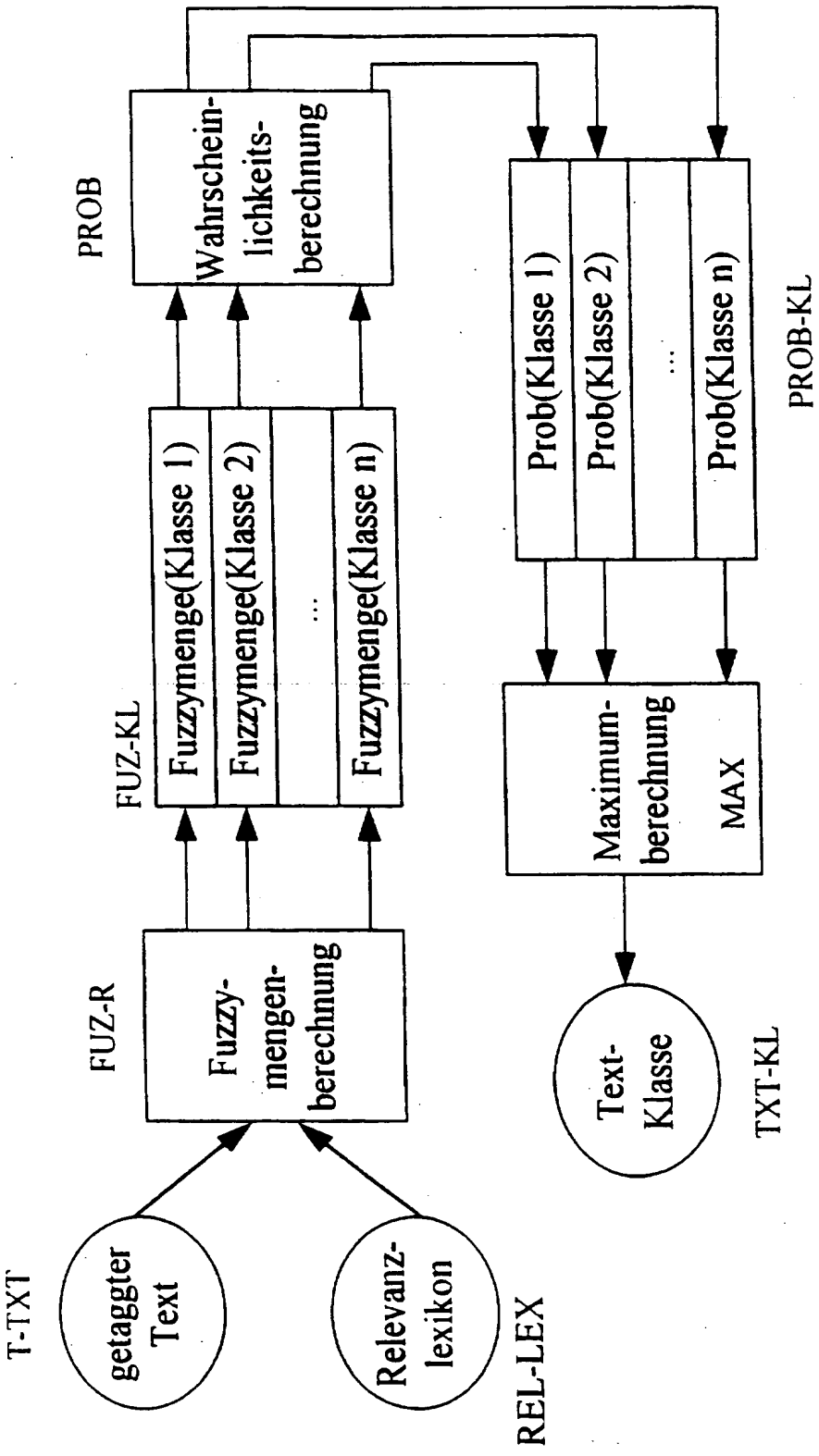
Figur 1



Figur 2


REL
Relevanz-
berechnung
Wort-
frequenzen
Klassen-
frequenzen
FR-W
FR-KL
FR
Frequenz-
berechnung
REL-LEX
Relevanz-
lexikon
TXT-TR
KL-TXT
Text 1
Text 2
...
Text n
Klasse(Text 1)
Klasse(Text 2)
...
Klasse(Text n)

Figur 3


T-TXT
getaggter Text
REL-LEX
Relevanz-lexikon
FUZ-R
Fuzzy-mengen-berechnung
FUZ-KL
Fuzzymenge(Klasse 1)
Fuzzymenge(Klasse 2)
...
Fuzzymenge(Klasse n)
PROB
Wahrschein-lichkeits-berechnung
Prob(Klasse 1)
Prob(Klasse 2)
...
Prob(Klasse n)
PROB-KL
Maximum-berechnung MAX
Text-Klasse
TXT-KL

Figur 4

INTERNATIONAL SEARCH REPORT

Inte[illegible]nal Application No
[illegible]T/DE 97/00583

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 G06F17/60 G06F17/30 G06K9/20

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 G06F G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	SIGIR '94. PROCEEDINGS OF THE SEVENTEENTH ANNUAL INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, PROCEEDINGS OF 17TH INTERNATIONAL CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL. SIGIR 94, DUB, ISBN 3-540-19889-X, 1994, BERLIN, GERMANY, SPRINGER-VERLAG, GERMANY, pages 31-40, XP000475312 HOCH R: "Using IR techniques for text classification in document analysis" cited in the application see the whole document --- -/--	1-8

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

1

Date of the actual completion of the international search	Date of mailing of the international search report
18 July 1997	29.07.97
Name and mailing address of the ISA European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+ 31-70) 340-2040, Tx. 31 651 epo nl, Fax (+ 31-70) 340-3016	Authorized officer Suendermann, R

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

International Application No
PCT/DE 97/00583

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	INTERNATIONAL ASSOCIATION FOR PATTERN RECOGNITION WORKSHOP ON DOCUMENT ANALYSIS SYSTEMS, PROCEEDINGS OF THE INTERNATIONAL ASSOCIATION FOR PATTERN RECOGNITION WORKSHOP, KAISERSLAUTERN, GERMANY, OCT. 1994, ISBN 981-02-2122-3, 1995, SINGAPORE, WORLD SCIENTIFIC, SINGAPORE, pages 52-75, XP002035594 DENGEL A ET AL: "OfficeMAID-a system for office mail analysis, interpretation and delivery" cited in the application 8. Message Classification	1-8
X	US 5 371 807 A (REGISTER MICHAEL S ET AL) 6 December 1994 see abstract; claims 1-24	1
X	US 5 463 773 A (SAKAKIBARA YASUBUMI ET AL) 31 October 1995 see abstract; claims 1-6	1
P,A	EP 0 704 810 A (HITACHI LTD) 3 April 1996 see abstract; claims 1-11; figures 1,8,16,18,24	1
A	US 5 255 187 A (SORENSEN MARK C) 19 October 1993 see abstract; claim 1	1,8
A	EP 0 515 714 A (SIEMENS AG) 2 December 1992 cited in the application see abstract; claims 1-5; figure 1	1

1

INTERNATIONAL SEARCH REPORT

International Application No
PCT/DE 97/00583

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5371807 A	06-12-94	NONE	
US 5463773 A	31-10-95	JP 5324726 A	07-12-93
EP 0704810 A	03-04-96	JP 8153121 A	11-06-96
US 5255187 A	19-10-93	CA 2039615 A	04-10-91
EP 0515714 A	02-12-92	NONE	

INTERNATIONALER RECHERCHENBERICHT

Internales Aktenzeichen
PCT/DE 97/00583

A. KLASSIFIZIERUNG DES ANMELDUNGSGEGENSTANDES
IPK 6 G06F17/60 G06F17/30 G06K9/20

Nach der Internationalen Patentklassifikation (IPK) oder nach der nationalen Klassifikation und der IPK

B. RECHERCHIERTE GEBIETE

Recherchierter Mindestprüfstoff (Klassifikationssystem und Klassifikationssymbole)
IPK 6 G06F G06K

Recherchierte aber nicht zum Mindestprüfstoff gehörende Veröffentlichungen, soweit diese unter die recherchierten Gebiete fallen

Während der internationalen Recherche konsultierte elektronische Datenbank (Name der Datenbank und evtl. verwendete Suchbegriffe)

C. ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie*	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
X	SIGIR '94. PROCEEDINGS OF THE SEVENTEENTH ANNUAL INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, PROCEEDINGS OF 17TH INTERNATIONAL CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL. SIGIR 94, DUB, ISBN 3-540-19889-X, 1994, BERLIN, GERMANY, SPRINGER-VERLAG, GERMANY, Seiten 31-40, XP000475312 HOCH R: "Using IR techniques for text classification in document analysis" in der Anmeldung erwähnt siehe das ganze Dokument --- -/--	1-8

[X] Weitere Veröffentlichungen sind der Fortsetzung von Feld C zu entnehmen

[X] Siehe Anhang Patentfamilie

* Besondere Kategorien von angegebenen Veröffentlichungen :

"A" Veröffentlichung, die den allgemeinen Stand der Technik definiert, aber nicht als besonders bedeutsam anzusehen ist

"E" älteres Dokument, das jedoch erst am oder nach dem internationalen Anmeldedatum veröffentlicht worden ist

"L" Veröffentlichung, die geeignet ist, einen Prioritätsanspruch zweifelhaft erscheinen zu lassen, oder durch die das Veröffentlichungsdatum einer anderen im Recherchenbericht genannten Veröffentlichung belegt werden soll oder die aus einem anderen besonderen Grund angegeben ist (wie ausgeführt)

"O" Veröffentlichung, die sich auf eine mündliche Offenbarung, eine Benutzung, eine Ausstellung oder andere Maßnahmen bezieht

"P" Veröffentlichung, die vor dem internationalen Anmeldedatum, aber nach dem beanspruchten Prioritätsdatum veröffentlicht worden ist

"T" Spätere Veröffentlichung, die nach dem internationalen Anmeldedatum oder dem Prioritätsdatum veröffentlicht worden ist und mit der Anmeldung nicht kollidiert, sondern nur zum Verständnis des der Erfindung zugrundeliegenden Prinzips oder der ihr zugrundeliegenden Theorie angegeben ist

"X" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann allein aufgrund dieser Veröffentlichung nicht als neu oder auf erfinderischer Tätigkeit beruhend betrachtet werden

"Y" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann nicht als auf erfinderischer Tätigkeit beruhend betrachtet werden, wenn die Veröffentlichung mit einer oder mehreren anderen Veröffentlichungen dieser Kategorie in Verbindung gebracht wird und diese Verbindung für einen Fachmann naheliegend ist

"&" Veröffentlichung, die Mitglied derselben Patentfamilie ist

1

Datum des Abschlusses der internationalen Recherche	Absendedatum des internationalen Recherchenberichts
18.Juli 1997	29.07.97

Name und Postanschrift der Internationale Recherchenbehörde	Bevollmächtigter Bediensteter
Europäisches Patentamt, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+ 31-70) 340-2040, Tx. 31 651 epo nl, Fax (+ 31-70) 340-3016	Suendermann, R

Formblatt PCT/ISA/210 (Blatt 2) (Juli 1992)

INTERNATIONALER RECHERCHENBERICHT

Internationales Aktenzeichen
DE 97/00583

C.(Fortsetzung) ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie*	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
X	INTERNATIONAL ASSOCIATION FOR PATTERN RECOGNITION WORKSHOP ON DOCUMENT ANALYSIS SYSTEMS, PROCEEDINGS OF THE INTERNATIONAL ASSOCIATION FOR PATTERN RECOGNITION WORKSHOP, KAISERSLAUTERN, GERMANY, OCT. 1994, ISBN 981-02-2122-3, 1995, SINGAPORE, WORLD SCIENTIFIC, SINGAPORE, Seiten 52-75, XP002035594 DENGEL A ET AL: "OfficeMAID-a system for office mail analysis, interpretation and delivery" in der Anmeldung erwähnt 8. Message Classification	1-8
X	US 5 371 807 A (REGISTER MICHAEL S ET AL) 6.Dezember 1994 siehe Zusammenfassung; Ansprüche 1-24	1
X	US 5 463 773 A (SAKAKIBARA YASUBUMI ET AL) 31.Oktober 1995 siehe Zusammenfassung; Ansprüche 1-6	1
P,A	EP 0 704 810 A (HITACHI LTD) 3.April 1996 siehe Zusammenfassung; Ansprüche 1-11; Abbildungen 1,8,16,18,24	1
A	US 5 255 187 A (SORENSEN MARK C) 19.Oktober 1993 siehe Zusammenfassung; Anspruch 1	1,8
A	EP 0 515 714 A (SIEMENS AG) 2.Dezember 1992 in der Anmeldung erwähnt siehe Zusammenfassung; Ansprüche 1-5; Abbildung 1	1

1

INTERNATIONALER RECHERCHENBERICHT

Internationales Aktenzeichen
PCT/DE 97/00583

Im Recherchenbericht angeführtes Patentdokument	Datum der Veröffentlichung	Mitglied(er) der Patentfamilie	Datum der Veröffentlichung
US 5371807 A	06-12-94	KEINE	
US 5463773 A	31-10-95	JP 5324726 A	07-12-93
EP 0704810 A	03-04-96	JP 8153121 A	11-06-96
US 5255187 A	19-10-93	CA 2039615 A	04-10-91
EP 0515714 A	02-12-92	KEINE	